

Visual Theory of Mind for Human-Agent Collaboration in Smart Environments

Sandy Hoffmann¹, Stefan Sarkadi², Alison R. Panisson¹

¹Federal University of Santa Catarina (UFSC) – Brazil

²University of Lincoln – United Kingdom

sandy.hoffmann@posgrad.ufsc.br, ssarkadi@lincoln.ac.uk,
alison.panisson@ufsc.br

Abstract. *Smart environments increasingly rely on intelligent agents to support people in daily activities, particularly in assisted living scenarios. For such systems to support collaboration with humans, agents must move beyond reactive perception and develop an explicit understanding of human states, intentions, and needs. This paper presents an approach for human-agent collaboration in smart environments based on Visual Theory of Mind (VToM), enabling an assistive agent to infer human intentions and contextual beliefs from visual inputs. These visual inferences are integrated with additional multimodal signals, including contextual and interaction-based inputs, to support proactive and adaptive assistance. By combining visual reasoning with multimodal information, the proposed approach allows the agent to align its behavior with human goals and ongoing activities, fostering effective collaboration rather than unilateral automation. We discuss the agent architecture, inference mechanisms, and collaborative interaction scenarios in smart homes and assisted living environments, highlighting how VToM contributes to human-centered assistance in shared environments.*

1. Introduction

Smart environments have evolved into AI-driven spaces that integrate heterogeneous sensing and communication technologies to support human activities in a proactive and adaptive manner [Varol et al. 2026]. In these settings, assistive agents play a central role by continuously perceiving the environment, reasoning about human behavior, and acting to provide context-aware support. Recent advances in artificial intelligence have enabled agents to move beyond reactive automation toward collaborative assistance, requiring mechanisms to interpret users’ intentions, beliefs, and ongoing activities from multimodal inputs [Rocha et al. 2023].

Despite this progress, many smart environments systems remain primarily reactive and request-driven, relying on explicit commands or predefined workflows [Polo-Rodríguez et al. 2025, Rivkin et al. 2024]. Even when enhanced with Large Language Models (LLMs), their reasoning is often weakly grounded in physical perception and lacks an explicit representation of the user’s mental state [Luo et al. 2024]. As a result, these systems struggle to anticipate needs or adapt their responses according to what the user knows, intends, or mistakenly believes.

Theory of Mind (ToM) provides a cognitive framework for attributing mental states such as beliefs and intentions to others [Premack and Woodruff 1978, Goldman 2012].

In smart environments, this reasoning must be grounded in visual context, motivating the adoption of Visual Theory of Mind (VToM) as a computational extension that infers mental states from observed scenes [Zhou et al. 2023, Hoffmann et al. 2025].

In this paper, we propose a proactive assistive agent architecture grounded in VToM. The approach integrates structured visual perception, a persistent world state, and LLM-based reasoning to infer user intentions and beliefs from contextual and visual cues. By explicitly modeling the user’s perspective and knowledge state, the agent provides belief-aware, context-sensitive assistance rather than purely reactive responses. The main contributions of this work are: (i) a VToM-based architecture that grounds mental-state inference in visual perception; (ii) a formulation of proactivity emerging from the interaction between world state, inferred intentions, and modeled beliefs; and (iii) an evaluation using multimodal ToM scenarios demonstrating intention-based and belief-aware assistance in smart environments.

2. Background

2.1. Smart Environments and AI Agents

Smart environments are systems capable of acquiring and applying knowledge about the environment and its users in order to improve their experience in that environment [Cook and Das 2007, Youngblood et al. 2005].

Smart environments are undergoing a significant transformation with the introduction of AI agents powered by multi-modal LLMs, shifting automation from predefined rule-based behaviors toward adaptive, context-aware collaboration. Automation can be understood as a continuous cycle of perceiving the state of the environment, reasoning about this state in conjunction with task goals and the expected outcomes of possible actions, and acting upon the environment [Cook and Das 2007]. AI agents naturally align with this paradigm, as they explicitly embody perception-reasoning-action loops and can integrate heterogeneous sensory inputs with high-level symbolic and semantic reasoning enabled by LLMs.

Rather than operating as isolated controllers, such agents maintain an internal representation of the environment and its occupants, allowing them to interpret evolving situations, understand users intentions, anticipate human needs, and adjust their behavior accordingly. This agent-based perspective supports a transition from reactive automation to collaborative assistance, in which intelligent systems participate in shared environments as adaptive, goal-aware partners.

2.2. Visual Theory of Mind (VToM)

Recent advances in distributed AI have highlighted the importance of autonomous agents capable of collaboration and adaptive reasoning, aligning with long-term efforts toward General Artificial Intelligence (GAI) [Goertzel 2014]. A core requirement for such systems is the ability to reason about other agents and humans in a socially aware manner, which is fundamentally associated with Theory of Mind (ToM) [Goldman 2012, Premack and Woodruff 1978]. ToM refers to the cognitive capacity to attribute mental states, such as beliefs, desires, and intentions, to oneself and others, enabling prediction of behavior, cooperation, and decision-making under uncertainty [Rocha et al. 2023, Chen et al. 2021].

As intelligent systems increasingly operate in perceptually rich and multimodal environments, traditional ToM approaches based primarily on symbolic or linguistic reasoning have shown limitations. Human cognition relies heavily on visual perception to infer others’ attention, intentions, and situational awareness, motivating the emergence of VToM as an extension of ToM grounded in visual cues [Zhou et al. 2023, Hoffmann et al. 2025, Hoffmann et al. 2026b]. VToM focuses on interpreting perceptual signals such as gaze direction, gestures, spatial relations, and observed actions to infer mental states directly from visual scenes. Computationally, VToM is commonly associated with the Belief-Desire-Intention (BDI) model [Bratman 1987] and the Simulation Theory of Mind (SToM), in which agents reason about others by internally simulating their perceptual and situational context [Goldman 2012].

Prior research has explored VToM-related mechanisms in applications such as robot perspective-taking [Johnson and Demiris 2007], shared attention modeling [Marin et al. 2008], intelligent monitoring systems [Hoffmann et al. 2026a], and behavioral validation tasks demonstrating human-like response patterns [Fischer and Demiris 2019]. More recently, advances in multimodal and vision-language models have renewed interest in grounding ToM-related reasoning in perceptual data, highlighting VToM as a key enabler for human-agent collaboration in smart environments [Hoffmann et al. 2025].

In this work, we investigate the application of VToM in smart environments, enabling assistive agents to infer users’ mental states, such as intentions and beliefs, and to provide proactive support during ongoing activities.

3. Related Work

Within the scope of this paper, the survey in [Varol et al. 2026] provides an overview of key technologies for intelligent assistive agents in smart environments, emphasizing multimodal perception, agent-based reasoning, and human-centered interaction. It highlights the role of visual understanding, semantic reasoning, and ToM-inspired models in enabling adaptive and collaborative assistance, positioning VToM as a relevant capability for next-generation assistive agents.

In [Polo-Rodríguez et al. 2025], the authors propose a context-aware smart environment architecture that integrates UWB-based user localization and human activity recognition with an LLM-powered chatbot to generate personalized, context-dependent interactions. While effective in adapting responses to real-time user context, this approach is primarily dialogue-driven and relies on explicit sensing infrastructure. In contrast, our work focuses on an assistive agent that directly observes users through visual inputs and proactively infers intentions and beliefs using VToM, without requiring explicit user queries or wearable localization devices.

In [Luo et al. 2024], the authors examine the broader challenge of integrating LLMs with the physical world, surveying techniques such as IoT sensing, reinforcement learning, embodied agent training, and robot task planning. Rather than addressing system-level integration in general, our approach concentrates on perceptual grounding through visual segmentation and semantic extraction, providing a structured description of the physical environment that enables VToM-based inference of user intentions and beliefs.

In [Rivkin et al. 2024], the authors introduce a smart home assistant that uses LLM-controlled action sequences to execute user requests via dynamic prompt trees, achieving strong performance on benchmark tasks. However, the system remains fundamentally request-driven. By contrast, our work emphasizes proactive assistance grounded in visual perception and Theory of Mind, enabling anticipation and collaboration rather than reactive execution of explicit commands.

4. A Proactive VToM-Based Assistive Agent for Smart Environments

This section introduces the architecture of the proposed VToM-Based assistive agent, which utilizes a processing pipeline to help users accomplish their goals, as illustrated in Figure 1. This work integrates computer vision with an LLM analysis within a dynamic framework to enable real-time responses, triggered by user observation/interaction. The proposed approach is designed to proactively support users in smart environments by continuously reasoning over multiple sources of information rather than reacting to explicit commands.

The framework is optimized for structured scenarios with stable spatial configurations. It is particularly effective in settings where fixed perspective camera are utilized, such as residential living spaces. By mapping the relationship between the user’s real-time location relative to known objects and containers, the system provides valuable data to LLM analysis. This allows the agent to offer navigation guidance through the user’s own point of view.

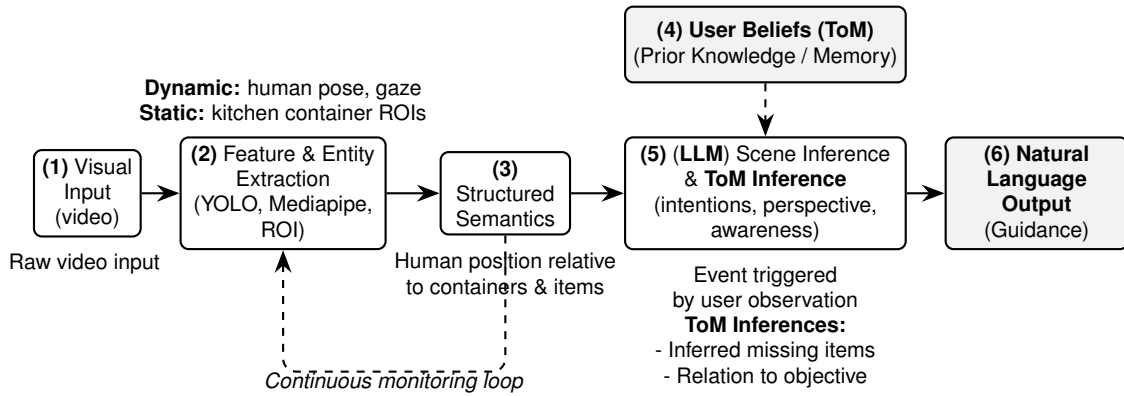


Figure 1. VToM-Based Assistive Agent Pipeline

4.1. Proposed Architecture

To enable the detection and mapping the user’s position, the system employs a combination of YOLO (*You Only Look Once*) [Jocher et al. 2023], and MediaPipe [Lugaresi et al. 2019] as part of the Visual Input (1) and Feature & Entity Extraction (2) stages. Once the regions of interest (ROI) for each container are mapped within the video, YOLO tracks the user’s movement by specifically targeting ‘human’ detections. MediaPipe complements this by enhancing the user’s positional data, utilizing pose references (such as left shoulder and right shoulder) as key parameters. In instances where the user’s body is partially occluded by objects, the MediaPipe face detection module serves as a failsafe to ensure continuous tracking.

This information is used to maintain a Structured Semantic (3) model of the current world state. It captures the user’s exact location alongside a list of mapped containers, which serves as a spatial reference for directional linguistics (e.g., left, right, or center) relative to the user. Each container maintains a record of its contents, tracking the specific objects stored within it.

The framework incorporates a set of User Beliefs (4), tracking their assumptions regarding object locations. The LLM Scene and ToM Inference (5) agent is designed to avoid repeating accurate users beliefs, focusing on correcting erroneous ones. When triggered, the agent generates a inference by processing the world state data alongside the inferred user goal and a detailed prompt instruction. With this, the LLM outputs the Natural Language Output (6) containing the required items to fulfill the user’s intention.

4.2. World State

To elaborate further on this component, the world state represents the agent’s internal model of the smart environment, capturing information about objects, spatial configurations, and relevant environmental conditions inferred from past and current observations. This representation is continuously updated as new perceptual data become available, allowing the agent to reason about changes in the environment over time.

By maintaining an explicit and persistent world state, the agent can ground its decisions in contextual knowledge, such as object locations and recent interactions, which is essential for reliable and proactive assistance. Listing 1 demonstrates an example of the world state representation for the scenario illustrated in Figure 2. The world state representation is designed to be modular, facilitating the integration of additional features. For example, identifying which objects the user is interacting with at a given moment. This information can be used to automatically infer user beliefs. For example, if a user opens a cabinet containing a specific item, the agent may infer that the user has observed its location. Consequently, this knowledge becomes part of the agent’s ToM about the user, modeling an updated belief state. This enables the LLM to filter its responses by avoiding references to items whose locations the user is already inferred to know.

4.3. Theory of Mind about the Users

The agent’s ToM models the user’s mental state, including beliefs, goals, and inferred intentions, and serves as the cognitive basis for human-centered assistance. This model is constructed by integrating multiple information sources, combining the proposed VToM component with interaction-based inputs obtained through textual or audio interfaces [da Silva et al. 2024]. While VToM provides perceptually grounded inferences from visual observations, additional interaction signals complement this understanding, enabling a richer and more robust representation of the user’s internal state.

The structured extraction pipeline enables the LLM model to infer user intent by maintaining a user-centric viewpoint. This process relies on integration the detailed world state into the model’s reasoning framework.

4.4. Proactivity by Inferring User’s Intentions

Proactive assistance emerges from the agent’s ability to infer triggering user intentions by jointly reasoning over the world state and the estimated mental state, including users’ goals expressed through communication interfaces.

Rather than responding to explicit commands, the agent identifies situations in which user behavior, environmental context, and previous interactions suggest an upcoming need or goal. This inference process enables the agent to decide when intervention is appropriate and what form of assistance should be provided, supporting timely and context-aware collaboration in smart environments.

5. Evaluation

To evaluate the proposed approach, we use a subset of scenarios from the Multimodal Theory of Mind Question Answering (MMToM-QA) benchmark [Jin et al. 2024]. The dataset consists of 134 videos recorded in a simulated household environment, depicting a person searching for objects. It includes 600 multiple-choice questions related to the person’s goals and beliefs, with two answer options per question. The code used for this experiment is available on a public GitHub¹ repository.

This work aims at a more comprehensive evaluation by using different scenario configurations. These scenarios provide representative situations for assessing the proposed assistive agent and its ability to infer users’ mental states. In particular, the inferred mental models are contrasted with the ground-truth mental states provided by the benchmark.

5.1. Scenario 1 – Assisting with Drinking Wine Based on User Intention

In the first scenario, the user intends to drink wine. This intention is inferred by the agent based on prior verbal interactions in which the user expresses a desire for wine, together with the observed scene shown in Figure 2, where the user enters the kitchen searching for an object.

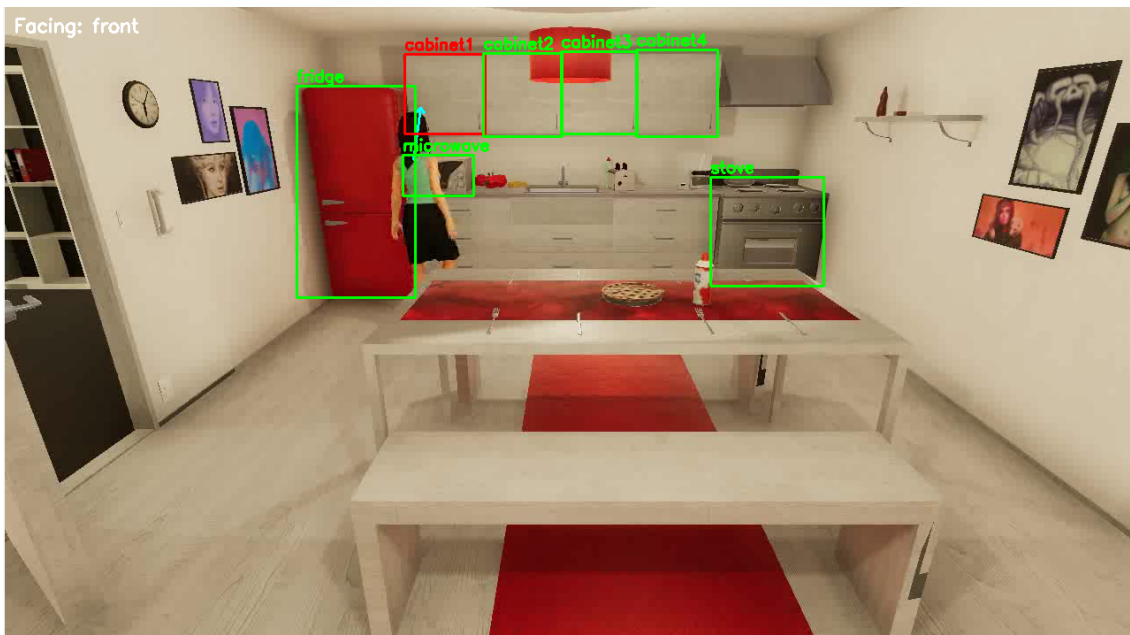


Figure 2. Scenario 1 - wine at cabinet 1 and wine glass at cabinet 3

In this scenario, no relevant user beliefs are considered. The inferred user intention is “The user intends to drink wine”, and the agent proactively assists the

¹https://github.com/SandyHoffmann/Visual_Tom_Experiment_YOLO

user by locating the wine and the necessary utensils for consumption. Listing 1 illustrates the world state representation for this scenario.

The agent output for this scenario is presented in Listing 2. It can be observed that the agent is capable of correctly inferring the necessary utensils required to drink the wine and successfully assisting the user in locating both the wine and the utensils, thereby providing complete assistance aligned with the user's intention².

Listing 1. Environment state representation

```
1 [World State]:
2 { "human": { "position_px": [471,172], "facing_direction": "front",
3   "looking_at": "cabinet1", "touching": null},
4   "objects": {
5     "fridge": { "contains": [ "apple", "salmon", "plate", "dish bowl
6       " ], "relative_to_human": "same depth and to the left" },
7     "cabinet1": { "contains": [ "wine", "wine glass", "bottle" ], "
8       relative_to_human": "in front and to the center" },
9     "cabinet2": { "contains": [], "relative_to_human": "in front and
10    to the right" },
11    "cabinet3": { "contains": [ "water glass", "dish bowl" ], "
12      relative_to_human": "in front and to the right" },
13    "cabinet4": { "contains": [], "relative_to_human": "in front and
14      to the right" },
15    "stove": { "contains": [ "cupcake" ], "relative_to_human": "
16      behind and to the right" },
17    "microwave": { "contains": [ "salmon" ], "relative_to_human": "
18      same depth and to the center" }}}
```

Listing 2. Agent output - scenario 1

```
1 The wine is in cabinet1, which is in front and to the center of your
2   current position, and there are a wine glass is in cabinet3,
3   which is in front and to the right of your current position.
```

5.2. Scenario 2 – Assisting with Eating a Sandwich Based on User Intention

In the second scenario, the user intends to eat a sandwich. As in the previous scenario, this intention is assumed to be inferred from prior interactions with the user, for example, through a verbal statement such as “I am very hungry; I think I will eat a sandwich.”

The scene corresponding to this scenario is illustrated in Figure 3. The world state is also depicted in Figure 3 through textual labels inside the bounding boxes, indicating, for instance, that the cheese is located inside the refrigerator.

For this scenario, the inferred user intention is “The user intends to eat a sandwich,” and no relevant user beliefs are considered. Given the current world state, which indicates that no prepared sandwich is available in the kitchen, the agent proactively assists the user by informing them of the locations of the available ingredients to prepare a sandwich. The agent output is presented in Listing 3. The agent is able not only to reason about the necessary ingredients but also to correctly identify and communicate the location of each item, thereby supporting the user in achieving their goal.

²Additional scenarios are presented in the Google Forms employed for the subjective evaluation of the proposed approach: <https://tinyurl.com/3f5my22m>

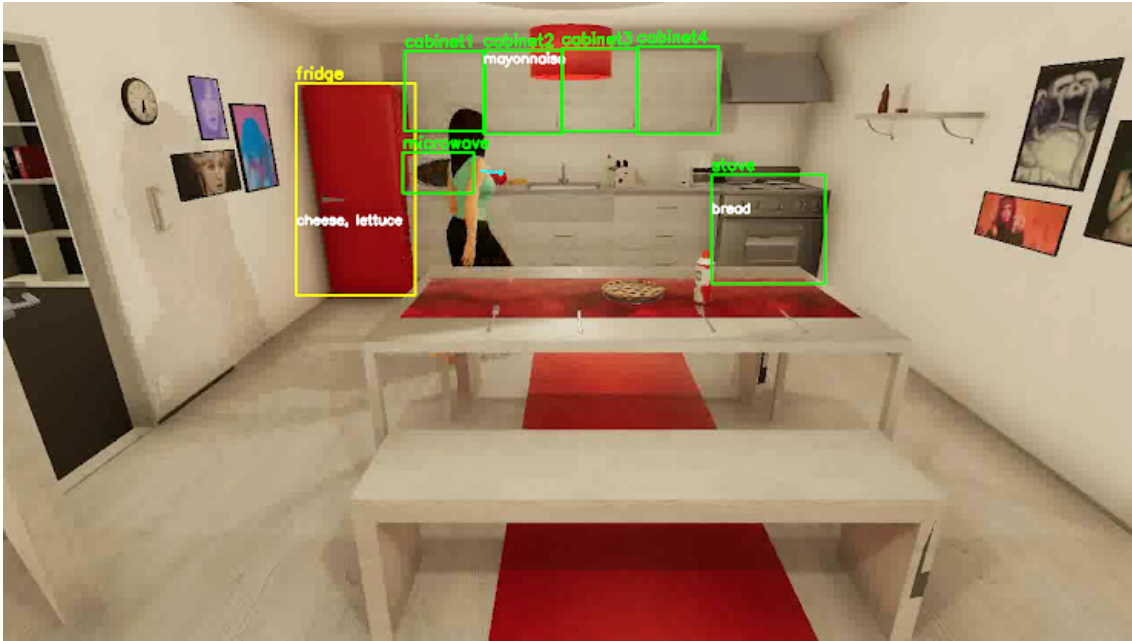


Figure 3. Scenario 2 - cheese and lettuce at fridge, mayonnaise at cabinet 2 and bread at stove.

Listing 3. Agent output - scenario 2

```

1 To make a sandwich, we'll need some ingredients. Based on the
  current world state, here's where we can find some common
  sandwich ingredients:
2 Lettuce: You can find lettuce inside the fridge, which is in front
  of you and to your left.
3 Cheese: Cheese is also located in the fridge, which is in front of
  you and to your left.
4 Bread: Bread is on the stove, which is in front of you and to your
  right.
5 Mayonnaise: Mayonnaise is in cabinet2, which is in front of you and
  in the center.
6 Let me know if you need help finding anything else or if you have
  other ingredients in mind!

```

5.3. Scenario 3 – Assisting with Making Pancakes Based on User Intention

In the third scenario, the user intends to make pancakes. As in the previous scenarios, this intention is inferred from prior user inputs, such as verbal or textual expressions of intent. The scene corresponding to this scenario is illustrated in Figure 4. The world state is also depicted in Figure 4 through textual labels, following the same representation used in the previous scenarios.

For this scenario, the inferred user intention is “The user intends to make pancakes,” and no relevant user beliefs are considered. Given the current world state, the agent proactively assists the user by informing them of the locations of the required ingredients. The agent output is presented in Listing 4. The agent is able not only to reason about the necessary ingredients but also to correctly identify and communicate the location of each item, including ingredients that are available in multiple locations.



Figure 4. Scenario 3 - milk at cabinet 1, butter in both fridge and microwave, flour, sugar, and salt at cabinet 3, and eggs at fridge

Listing 4. Agent output - scenario 3

```

1 Here's where you can find the ingredients for pancakes :
2 Flour: You can find the flour in cabinet3, which is in front and to
   your right.
3 Milk: The milk is in cabinet1, which is in front and to the center.
4 Eggs: The eggs are in the fridge, which is in front and to your
   right.
5 Sugar: You can find the sugar in cabinet3, which is in front and to
   your right.
6 Salt: The salt is in cabinet3, which is in front and to your right.
7 Butter: There's butter in the fridge (in front and to your right)
   and also in the microwave (in front and to the center).

```

5.4. Scenario 4 – Assisting with Drinking Wine Based on User Intention and Theory of Mind

In this scenario, the user intends to drink wine, similarly to Scenario 1 in Section 5.1. However, unlike the previous scenarios, the agent explicitly considers its ToM model, incorporating the user's beliefs into the reasoning process. In this particular case, illustrated in Figure 5, the agent's ToM includes the belief `believes(user at(cabinet1, wine))`, indicating that the agent knows that the user is already aware that the wine is located in `cabinet1`. The agent output is presented in Listing 5. Based on this belief, the agent provides only the information necessary for the user to achieve their goal, deliberately omitting information that the user is inferred to already know. This behavior illustrates how ToM enables more concise, personalized, and context-aware assistance.



Figure 5. Scenario 4 - wine is at cabinet 1 and wine glass at cabinet 3

Listing 5. Agent output - scenario 3

```
1 To drink wine, you'll need a wine glass. You can find a wine glass
  in cabinet3, which is in front and to your left.
```

5.5. Scenario 5 – Assisting with Making Pancakes Based on User Intention and Theory of Mind

In this scenario, similarly to the scenario described in Section 5.3, the user intends to make pancakes, and the assistive agent reasons about the user's beliefs regarding the locations of some ingredients. The agent's ToM includes the beliefs `believes(user, at(cabinet3, eggs))`, indicating that the agent believes the user thinks the eggs are located in `cabinet3`, and `believes(user, at(microwave, butter))`, indicating that the agent believes the user thinks the butter is located in the microwave. However, the user holds a false belief, as evidenced in Figure 6, where the eggs are actually located in the refrigerator.

The agent output for this scenario is presented in Listing 6. In this case, the agent provides the locations of all ingredients that it infers the user does not correctly know, including those for which the user's beliefs about their locations are incorrect. This behavior illustrates the agent's ability to identify and correct false beliefs when providing assistance.

Listing 6. Agent output - scenario 3

```
1 To make pancakes, you'll need the following ingredients:
2 Flour is in cabinet3, which is in front and to your left.
3 Eggs are in the fridge, which is in front and to your right.
4 Milk is in cabinet1, which is in front and to your left.
5 Sugar is in cabinet3, which is in front and to your left.
6 Salt is in cabinet3, which is in front and to your left.
```



Figure 6. Scenario 5 - milk at cabinet 1, butter in both fridge and microwave, flour, sugar, and salt at cabinet 3, and eggs at fridge

5.6. Qualitative Analysis of Agent Outputs

To evaluate the agent’s output across different situations, we asked computer science and engineering professionals and students to complete an online questionnaire³. The instrument presented 11 representative scenarios, each composed of an input image, the inferred intention, the agent’s ToM interpretation (for some scenarios), and the generated response. Fourteen independent evaluators participated in the study and assessed the effectiveness of the agent’s assistance for each scenario using a 5-point Likert scale [Likert 1932, Krosnick 1999], ranging from “Very Poor” to “Excellent/Very Good”. The aggregated results provide a qualitative assessment of the perceived adequacy and usefulness of the agent’s reasoning and responses.

As illustrated in Figure 7, the overall distribution is strongly skewed toward positive evaluations. Across scenarios, the proportion of “Good” and “Excellent/Very Good” ratings dominates, exceeding 70% and reaching near-total agreement in some cases (e.g., Scenarios 1, 2, 7, and 8). Negative evaluations (“Very Poor” and “Poor”) remain limited and concentrated in a small subset of scenarios, most notably Scenario 5, which presents a comparatively higher proportion of critical assessments. Neutral responses (“Fair/Average”) appear in moderate proportions in specific cases (e.g., Scenarios 3, 4, and 6), suggesting partial uncertainty rather than clear dissatisfaction.

Overall, the results indicate that evaluators perceived the agent’s assistance as effective and contextually appropriate in the majority of situations. The predominance of positive ratings and the relatively low incidence of negative feedback suggest that the underlying reasoning and response-generation mechanisms operate reliably across heterogeneous contexts. When critical scores occur, they appear scenario-specific rather than systematic, pointing to potential edge cases where contextual interpretation or response calibration may be further refined rather than indicating structural limitations of the pro-

³The questionnaire is anonymously available at <https://tinyurl.com/3f5my22m>.

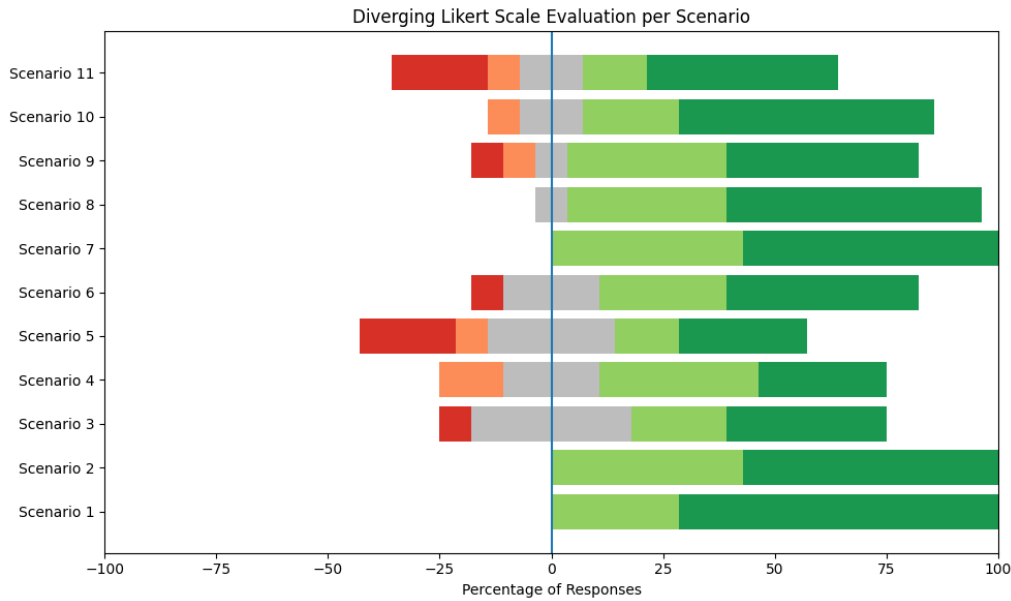


Figure 7. Distribution of answers based on the 5-point Likert Scale.

posed approach.

6. Conclusion

This paper presented a proactive assistive agent for smart environments grounded in VToM, enabling human-agent collaboration through perceptually informed reasoning. By integrating visual inference with multimodal interaction signals, the proposed approach allows the agent to infer users' intentions and beliefs, maintain an explicit world state, and provide context-aware assistance without relying solely on explicit commands.

The evaluated scenarios demonstrate how VToM supports increasingly sophisticated forms of assistance, ranging from intention-based guidance to belief-aware and false-belief-sensitive interaction. These results highlight the role of VToM as a key mechanism for shifting smart environment automation from reactive execution toward anticipatory, human-centered collaboration.

Future work will focus on more comprehensive evaluations in real-world settings, richer belief modeling, and tighter integration with embodied perception and interaction to further advance collaborative assistive agents in smart environments.

References

- Bratman, M. (1987). *Intention, Plans, and Practical Reason*. MA: Harvard University Press, Cambridge.
- Chen, B., Vondrick, C., and Lipson, H. (2021). Visual behavior modelling for robotic theory of mind. *Scientific Reports*, 11(1):424.
- Cook, D. J. and Das, S. K. (2007). How smart are our environments? an updated look at the state of the art. *Pervasive and mobile computing*, 3(2):53–73.
- da Silva, H. H., Rocha, M., Trajano, G., Morales, A. S., Sarkadi, S., and Panisson, A. R. (2024). Distributed theory of mind in multi-agent systems. In Rocha, A. P., Steels,

- L., and van den Herik, H. J., editors, *Proceedings of the 16th International Conference on Agents and Artificial Intelligence, ICAART 2024, Volume 1, Rome, Italy, February 24-26, 2024*, pages 451–460. SCITEPRESS.
- Fischer, T. and Demiris, Y. (2019). Computational modeling of embodied visual perspective taking. *IEEE Transactions on Cognitive and Developmental Systems*, 12(4):723–732.
- Goertzel, B. (2014). Artificial general intelligence: Concept, state of the art, and future prospects. *Journal of Artificial General Intelligence*, 5(1):1.
- Goldman, A. I. (2012). *Theory of Mind*. Oxford University Press, United Kingdom.
- Hoffmann, S., Fernandes, A. R., Muchulski, V. W., Sarkadi, S., von Wangenheim, A., and Panisson, A. R. (2026a). Towards Intelligent Monitoring System Using Computer Vision. In *In Proceedings of the International Conference on Agents and Artificial Intelligence (ICAART)*.
- Hoffmann, S., Trajano, G., Sarkadi, S., and Panisson, A. R. (2025). Visual Theory of Mind (VToM): A Systematic Review. In *IEEE/WIC International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*.
- Hoffmann, S., Vieira, L. N., Sarkadi, S., and Panisson, A. R. (2026b). Visual theory of mind through llm-based semantic extraction. *ToM4AI 2026*, page 118.
- Jin, C., Wu, Y., Cao, J., Xiang, J., Kuo, Y.-L., Hu, Z., Ullman, T. D., Torralba, A., Tenenbaum, J. B., and Shu, T. (2024). Mmtom-qa: Multimodal theory of mind question answering. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 1:16077 – 16102.
- Jocher, G., Chaurasia, A., and Qiu, J. (2023). Ultralytics yolov8.
- Johnson, M. and Demiris, Y. (2007). Visuo-cognitive perspective taking for action recognition. In *Int. Symp. Imitation Animals Artifacts*, pages 262–269.
- Krosnick, J. A. (1999). Maximizing questionnaire quality. *Measures of political attitudes*, 2:37–58.
- Likert, R. (1932). A technique for the measurement of attitudes. *Archives of psychology*.
- Lugaresi, C., Tang, J., Nash, H., McClanahan, C., Uboweja, E., Hays, M., Zhang, F., Chang, C.-L., Yong, M., Lee, J., Chang, W.-T., Hua, W., Georg, M., and Grundmann, M. (2019). Mediapipe: A framework for perceiving and processing reality. In *Third Workshop on Computer Vision for AR/VR at IEEE Computer Vision and Pattern Recognition (CVPR) 2019*.
- Luo, X., Liu, D., Dang, F., and Luo, H. (2024). Integration of llms and the physical world: Research and application. In *Proceedings of the ACM Turing Award Celebration Conference-China 2024*, pages 1–5.
- Marin, L., Sisbot, E. A., and Alami, R. (2008). Geometric tools for perspective taking for human-robot interaction. In *Mexican international conference on artificial intelligence (MICA I 2008), Mexico City, Mexico*.

- Polo-Rodríguez, A., Fiorini, L., Rovini, E., Cavallo, F., and Medina-Quero, J. (2025). Enhancing smart environments with context-aware chatbots using large language models. *arXiv preprint arXiv:2502.14469*.
- Premack, D. and Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and brain sciences*, 1(4):515–526.
- Rivkin, D., Hogan, F., Feriani, A., Konar, A., Sigal, A., Liu, X., and Dudek, G. (2024). Aiot smart home via autonomous llm agents. *IEEE Internet of Things Journal*.
- Rocha, M., da Silva, H. H., Morales, A. S., Sarkadi, S., and Panisson, A. R. (2023). Applying theory of mind to multi-agent systems: A systematic review. In *Brazilian Conference on Intelligent Systems*, pages 367–381. Springer.
- Varol, A., Motlagh, N. H., Leino, M., Tarkoma, S., and Virkki, J. (2026). Creation of ai-driven smart spaces for enhanced indoor environments—a survey. *Internet of Things*, page 101876.
- Youngblood, G. M., Heierman, E. O., Holder, L. B., and Cook, D. J. (2005). Automation intelligence for the smart environment. In *International Joint Conference On Artificial Intelligence*, volume 19, page 1513. LAWRENCE ERLBAUM ASSOCIATES LTD.
- Zhou, P., Ma, H., Zou, B., Zhang, X., Zhao, S., Lin, Y., Wang, Y., Feng, L., and Wang, G. (2023). A conceptual framework of cognitive-affective theory of mind: towards a precision identification of mental disorders. *npj Mental Health Research*, 2(1):12.